

Analyzing Lifestyle and Environmental Factors on Semen Fertility using Association Rule Mining

M. A. Anwar (Corresponding author)
College of Engineering and Computing
Al Ghurair University, Dubai Academic City
PO Box 37374, Dubai, United Arab Emirates
Tel: + 971-4-4200223 ext. 345 E-mail: anwar@agu.ac.ae

Naseer Ahmed
Institutional Effectiveness and Planning
Al Ghurair University, Dubai Academic City
PO Box 37374, Dubai, United Arab Emirates
Tel: + 971-4-4200223 ext. 291 E-mail: naseer@agu.ac.ae

Abstract

The data mining has been used to extract hidden knowledge more effectively for analysis of business, academic, agricultural, as well as medical data in contrast to the predefined queries or reports. This paper presents the impacts of lifestyle and environmental factors of a man on the fertility and quality of semen using association rule mining. The association rules have been mined from data collected by a normalized questionnaire from young volunteers and are found to be useful in predicting the quality of semen based on individual's lifestyle and environmental factors.

Keywords: Association rules, Knowledge Discovery, Fertility potential, Rule confidence

1. Introduction

The past several decades have witnessed a fast growth in the usage of Artificial Intelligence and knowledge mining as a means by which useful hidden information can be extracted to make informed decisions. The medical occupation is also no exception. The publication of a meta-analysis directed by Elisabeth Carlsen has generated a debate about possible decline in the semen quality (E. Carlsen, A. Giwercman, N. Keiding, and N. E. Skakkebaek, 1992). Various studies have suggested the effects of environmental or occupational factors (Giwercman & Giwercman, 2011; Wong, Zielhuis, Thomas, Merkus, & Steegers-Theunissen, 2003) as well as a certain lifestyle (Martini et al., 2004; Agarwal, Desai, Ruffoli, & Carpi, 2008) on the fertility of the semen.

The clinicians assess the male partner's fertility by using the data obtained from semen analysis (Kolettis, 2003) and comparing the obtained results with the corresponding reference value established by World Health Organization (WHO, 1999). The analysis of semen is a good forecaster of the male fertility potential (Bonde et al.,1998; Guzick et al., 2001; Slama et al., 2002), and is also necessary to evaluate the appropriateness of the candidates to become semen donors (Barratt et al., 1998; Carrell, Cartmill, Jones, Hatasaka, & Peterson, 2002). There is a high variability in the testicular function of an individual (Keel, 2006), so it is recommended to interpret the semen analysis results taking into account certain other factors i.e., fever, toxic exposure that could potentially modify the semen parameters (Rowe & Comhaire, 2000).

Data mining (also called knowledge discovery) is the method of analyzing data from different perspectives to discover interesting and useful information. The information gained through data mining has been effectively used in various sectors ranging from finance, agriculture to health and education. There are many data mining tools (Weka 2012), (XLMiner 2013), (KNIME 2013) available that allow users to analyze data from various aspects, categorize it, and discover the identified relationships. Technically, data mining is a technique of finding correlations or patterns among many fields in large databases. The data mining is fast becoming an interesting research area in healthcare discipline as it allows researcher to extract useful, previously unknown patterns from the medical datasets for better understanding, improved performance and assessment of the treatment process.

In this paper, assuming the impact of environmental factors and life habits on the semen quality, we use Apriori algorithm for association rule mining that can help in the evaluation of male fertility potential. The remaining part of the paper is organized as follows: section 2 discusses the knowledge discovery process and data collection in detail, section 3 presents the results and analysis, and in section 4 conclusion is drawn.

2. Knowledge Discovery Process and Data Collection

Knowledge Discovery (KD) process is one of the basic concepts of the field of Knowledge Discovery and Data mining (KDD). Figure 1 illustrates the knowledge discovery employed in the present study that we have adapted from (Anna Katerina Domminguez et al., 2010). Solid-line arrows represent important data processing steps leading towards the knowledge discovery whereas dotted-line arrows show the steps that may form an iterative cycle in order to refine the knowledge discovery process.

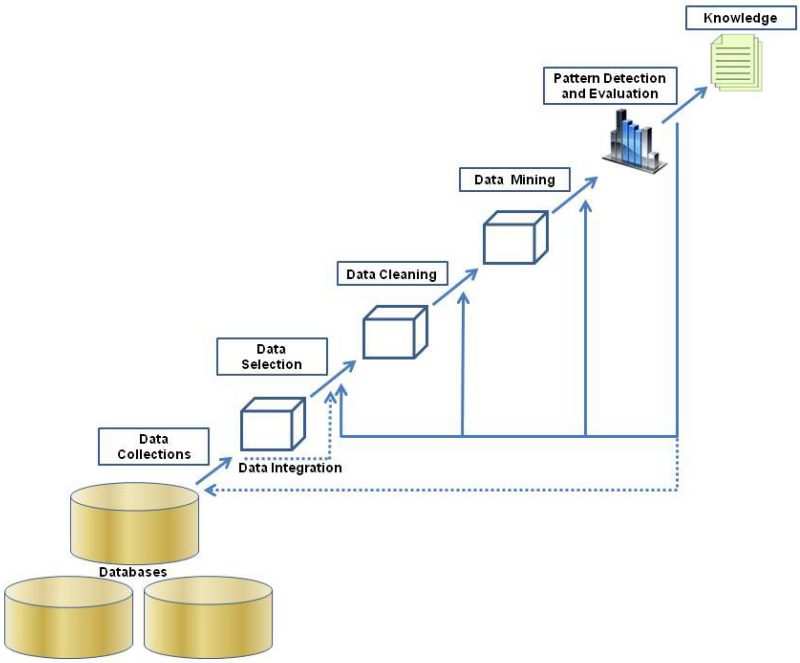

Figure 1. KDD Process.

2.1 Mining Frequent Patterns and Associations

The association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules show attributes' value conditions that occur frequently together in a given dataset (Han, J. et al., 2011). The preliminaries necessary for performing data mining on any data are discussed below.

Let $I = \{I_1, I_2, I_3, ..., I_m\}$ be a set items. Let D, the task relevant data, be a set of database transactions where each transaction $T \subseteq I$. Each transaction is an association with an identifier, called transaction identification (TID). Let A be a set of items. A transaction T is said to contain A if and only if $A \subseteq T$. An association rule is an implication of the form $A \Rightarrow B$, where $A \subset I, B \subset I$, and $A \cap B = \phi$.

Support (s) and confidence (c) are two measures of rule interestingness. They respectively reflect the usefulness and certainty of the discovered rule. A support of 2% of the rule $A \Rightarrow B$ means that A and B exist together in 2% of all the transactions under analysis. The rule $A \Rightarrow B$ having confidence of 60% in the transaction set D means that the percentage of transactions in D containing A that also contains B is 60.

A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. If the relative support of an itemset I satisfies a prescribed minimum support threshold, then I is a frequent itemset.

The association rule mining can be viewed as a two-step process:

1. Find all frequent itemsets: Each of these itemsets will occur at least as frequently as a predetermined minimum support count.

2. Generate strong association rules from the frequent itemsets: The rules must satisfy minimum support and confidence. These rules are called strong rules.

2.1 Apriori Algorithm

Apriori is a seminal algorithm proposed by (R. Agarwal et al., 1994) for mining frequent itemsets for Boolean association rules. The name of the algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties. The following lines state the steps in generating frequent itemset in Apriori algorithm.

Let *Ck* be a candidate itemset of size *k* and *Lk* as a frequent itemset of size *k*. The main steps of iteration are:

Find frequent set L_{k-1}

Join step: C_k is generated by joining L_{k-1} with itself (cartesian product $L_{k-1} \times L_{k-1}$)

Prune step (apriori property): Any (k − 1) size itemset that is not frequent cannot be a subset of a frequent k size itemset, hence should be removed

Frequent set L_k has been achieved

2.2 Task Relevant Data Collection

The data used in this study is available at UCI website (Bache, K. & Lichman, M., 2013) and is based on the semen analysis of 100 young volunteer donors between the age of 18 and 36 years. The samples of these volunteer donors were collected after 3 to 6 days of sexual abstinence and a semen analysis was performed according to the World Health Organization guidelines and standards. The donors having earlier reproductive alterations were excluded from the statistical analysis.

2.3 Data Preprocessing

The real-world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their likely origin from multiple, heterogeneous sources. Low-quality data will lead to low-quality mining results (Han, J. et al., 2011). Therefore, data preprocessing is an important task in data mining. The data we used in this paper is given in Table 1 which shows the name, a description and the range of values of the variables used in the study as well as the values normalized. The data set characteristics are multivariate and attributes are real in characteristics and for association rule mining the data must be categorized. Table 1 shows the preprocessing done to the different database input fields and the conversion of the input data into a range of normalization and categorization according to the following rules:

(a) Numerical variables such as age or frequency of alcohol consumption per day or per week are normalized onto interval (0, 1). For example the age at the time of analysis is shown in second column (Values) as a range between the minimum 18 and the maximum 36. This means that the donor having 36 years age is normalized to the value 36/36 = 1 whereas a donor with age of 27 is normalized to the value 27/36 = 0.75.

(b) The variables having only two independent attributes are prearranged with binary values 0 and 1.

(c) The variables with three independent attributes, such as ‘‘High fevers in the last year’’ and ‘‘Smoking habit’’ are prearranged using the ternary values (-1 , 0, 1). For instance, “High fever” will take -1 for never and 0 for less than three month ago and 1 for greater than three months ago.

(d) The variables with four independent attributes are prearranged using the four different and equal distance values (-1,-0.33, 0.33, 1).

The variables with a range of more than four independent attributes are considered numerical variables.

Table 1. Description of variable with their values range

Variable Description	Values (min-max)	Normalized	Categorized
Season in which the analysis was performed	1. Winter 2. Spring 3. Summer 4. Fall	(-1, -0.33, 0.33, 1)	Winter Spring Summer Fall
Age at the time of analysis.	18 – 36	(0, 1)	EAge (18 – 22) AAge (23 – 27) MAge (27 – 31) UAge (32 – 36)
Childish diseases (i.e. chicken pox, measles, mumps, polio)	1. Yes 2. No	(0, 1)	CDY (Child Disease Yes) CDN (Child Disease No)
Accident or serious trauma	1. Yes 2. No	(0, 1)	TY (Accident Yes) TN (Accident No)
Surgical intervention	1. Yes 2. No	(0, 1)	TY (Surgery Yes) TN (Surgery No)
High fevers in the last year	1. Less than three months ago 2. More than three months ago 3. No	(-1, 0, 1)	HFY (High Fever Yes) HFN (High Fever No)
Frequency of alcohol consumption	1. Several times a day 2. Every day 3. Several times a week 4. Once a week 5. Hardly ever or never	(0, 1)	AHigh AMedium ALow
Smoking habit	1. Never 2. Occasional 3. Daily	(-1, 0, 1)	Y (Yes) O (Occasional) N (N0)
Number of hours spent sitting per day	1 – 16	(0, 1)	SLong SMedium SShort
Output: Diagnosis	Normal (N), Altered (O)		Altered Normal

Table 2. Transformed fertility data

Data on Notepad File	Data Transformed into Categories									
	Weather	Age	Childish Disease	Trauma	Surgery	High Fever	Alcohol	Smoking	Sitting	Diagnosis
-0.33,0.69,0,1,1,0,0.8,0,0.88,N -0.33,0.94,1,0,1,0,0.8,1,0.31,O -0.33,0.5,1,0,0,0,1,-1,0.5,N -0.33,0.75,0,1,1,0,1,-1,0.38,N -0.33,0.67,1,1,0,0,0.8,-1,0.5,O -0.33,0.67,1,0,1,0,0.8,0,0.5,N -0.33,0.67,0,0,0,-1,0.8,-1,0.44,N -0.33,1,1,1,1,0,0.6,-1,0.38,N 1,0.64,0,0,1,0,0.8,-1,0.25,N 1,0.61,1,0,0,0,1,-1,0.25,N 1,0.67,1,1,0,-1,0.8,0,0.31,N	Spring	Uage	CDY	TN	SY	HFN	AHigh	Y	SShort	Altered
	Spring	Uage	CDY	TY	SN	HFN	AHigh	N	SMedium	Altered
	Fall	Uage	CDY	TN	SY	HFN	AHigh	N	SShort	Altered
	Fall	Uage	CDY	TN	SN	HFN	AHigh	Y	SMedium	Altered
	Fall	Uage	CDY	TN	SY	HFY	AHigh	N	SMedium	Altered
	Fall	Uage	CDY	TN	SY	HFN	AMedium	N	SMedium	Altered
	Fall	Uage	CDY	TY	SN	HFY	AMedium	N	SMedium	Altered
	Fall	Uage	CDN	TN	SY	HFN	AMedium	O	SMedium	Altered

2.3 Data Cleaning

It is a well-known fact that incorrect or inconsistent data can lead to false conclusions and hence wrong inferences and decisions. Therefore, high quality data needs to pass a set of quality criteria; accuracy, integrity, completeness, validity, consistency, uniformity, density, and uniqueness. Data cleaning routines attempts to fill in missing values, smooth out noise, and correct inconsistencies in the data. There are a number of data cleaning techniques (Han, J. et al., 2011) in the literature such as fill missing values, binning, regression, and clustering. We used the following criteria to clean the fertility data used in this paper:

(a) If a donor did not respond to any one of the items in the questionnaire then such record is removed from the dataset.

(b) If a donor has attempted many options in questionnaire then such record has also been removed from the dataset.

The raw data was on notepad file in "comma separated" format which was processed and transformed into categories to apply the association rule mining. The intermediate steps are not presented here. The sample raw data and pre-processed form of fertility data is show in Table 2.

3. Results and Rules Analysis

The association rules mined from fertility data are presented in Table 3, 4, and 5 with different supports and confidences. Due to the limitation of the space, a number of uninteresting rules have been excluded from Tables 3 – 5. The association rules depicted in these tables are mined using a data mining tool (XLMiner 2013). This tool allows mining the association rules by setting various minimum support thresholds. It is observed that by lowering the minimum support threshold there is a marked increase in the number of association rules generated by XLNimer tool. There are 69 association rules mined with minimum support 40 and minimum confidence 60%. We are interested in rules where consequent i.e. right hand side of the rule is either normal or altered. There are 21 such rules which show that whatever antecedent i.e. left side of the rule is, the fertility is not affected. For example, rule 6 with confidence of 97.62% shows that habit of high alcohol consumption does not affect the fertility of the sperm of donors between ages 18 – 36. Similarly rule 26 with confidence 88.41% shows that habit of high alcohol consumption and any disease in childhood also does not affect the fertility of the sperm. We could not find any rule at this low level of support and confidence showing that the behaviors listed in Table 1 affects the male fertility between ages of 18 – 36.

Table 3. Association rules mined; minimum support 40 and confidence 60%

Rule No.	Antecedent (a)		Consequent (c)	Support (a)	Support (c)	Support (a u c)	Confidence (%)
1	AHigh, CDY	⇒	Normal	69	88	61	87.5
2	AHigh, HFN	⇒	Normal	48	88	43	75
3	AHigh, Mage	⇒	Normal	42	88	41	88.41
4	AHigh, N	⇒	Normal	47	88	43	89.58
5	AHigh, TN	⇒	Normal	47	88	41	97.62
6	AHigh	⇒	Normal	79	88	71	91.49
7	CDY, HFN	⇒	Normal	53	88	47	87.23
8	CDY, N	⇒	Normal	48	88	42	89.87
9	CDY, SN	⇒	Normal	45	88	40	88.68
10	CDY	⇒	Normal	87	88	77	87.5
11	HFN	⇒	Normal	63	88	55	88.89
12	Mage	⇒	Normal	46	88	45	88.51
13	N	⇒	Normal	56	88	50	87.3
14	SMedium	⇒	Normal	51	88	43	97.83
15	SN	⇒	Normal	49	88	44	89.29
16	SY	⇒	Normal	51	88	44	84.31
17	TN	⇒	Normal	56	88	47	89.8
18	TY	⇒	Normal	44	88	41	86.27
19	Uage	⇒	Normal	69	88	61	83.93
20	AHigh, CDY	⇒	Normal	48	88	43	93.18
21	AHigh, HFN	⇒	Normal	42	88	41	88.41

Table 4: Association rules mined; minimum support 50 and confidence 70%

Rule No.	Antecedent (a)		Consequent (c)	Support (a)	Support (c)	Support (a u c)	Confidence (%)
1	N	⇒	Normal	56	88	50	89.29
2	CDY	⇒	Normal	87	88	77	88.51
3	AHigh, CDY	⇒	Normal	69	88	61	88.41
4	HFN	⇒	Normal	63	88	55	87.3

Table 5: Association rules mined; minimum support 70 and confidence 80%

Rule No.	Antecedent (a)		Consequent (c)	Support (a)	Support (c)	Support (a u c)	Confidence (%)
1	AHigh	⇒	Normal	79	88	71	89.87
2	CDY	⇒	Normal	87	88	77	88.51

There are 21 association rules mined with minimum support 40 and minimum confidence 60% as shown in Table 3. There are only four (04) such rules which show that whatever antecedent is, the fertility is not affected. For example, the rule number 15 with confidence of 89.29% shows that non-smokers fertility is always normal and similarly disease in childhood alone also does not affect the fertility.

There are four (04) association rules mined with minimum support 50 and minimum confidence 70% as shown in Table 4. There are only two (02) such rules (rule number 2 and 3) which show that high alcohol consumption or disease in childhood does not affect the fertility.

Table 5 shows the strongest rules with support 70 and confidence 80% that high alcohol and childhood disease do not affect the fertility.

We also mined rules with minimum support 10 and minimum confidence 50%. The total number of rules mined is 9623 and none of the rules consequent is altered. Therefore, we can conclude that the habits and behavior (listed in Table 1) of male between 18 -36 years age do not affect the fertility of sperm of a man.

4. Conclusion

The paper presented the potential use of one of the data mining approaches called association rule mining algorithm in finding the effect of human behavior and habit on the fertility of sperm of a man between the ages 18 – 36. The analysis reveals that variables described in Table 1 do not affect the fertility. However, it does not imply that these habits may be adopted. In future we plan to include more variable in the analysis and also use different data mining techniques to further find the hidden patterns in behavior and habits and fertility.

5. Acknowledgements

The authors wish to acknowledge the financial support provided by the Al Ghurair University and David Gil et. al. for providing the dataset available at UCI website.

References

Agarwal, A., Desai, N. R., Ruffoli, R., & Carpi, A. (2008), Biomed Pharmacother. 62(8):550-3.
Anna Katerina Domminguez, Kalina Yasef, and James R. Curran (2010). “Data Mining for Individualized Hints in eLearning”, in proceedings of EDM Educational Data Mining Conference, Pittsburg PA, USA.
Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository
[http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
Barratt, C. L., Clements, S., & Kessopoulou, E. (1998). Semen characteristics and fertility tests required for storage of spermatozoa. Human Reproduction (Oxford, England), 13(Suppl. 2), 1–7. discussion 8–11.

Bonde, J. P., Ernst, E., Jensen, T. K., Hjollund, N. H., Kolstad, H., Henriksen, T. B., et al. (1998). Relation between semen quality and fertility: A population-based study of 430 first-pregnancy planners. Lancet, 352(9135), 1172–1177.
Carrell, D. T., Cartmill, D., Jones, K. P., Hatasaka, H. H., & Peterson, C. M. (2002). Prospective, randomized, blinded evaluation of donor semen quality provided by seven commercial sperm banks. Fertility and Sterility, 78(1), 16–21.
David H. Shanabrook, David G. Cooper, Beverly Park Woolf, and Ivan Arroyo (2010). "Identifying High-Level Student Behavior Using Sequence-based Motif Discovery", in proceedings of EDM Educational Data Mining Conference, Pittsburg PA, USA.
E. Carlsen, A. Giwercman, N. Keiding, and N. E. Skakkebaek, BMJ. 1992 September 12; 305(6854): 609–613.
Giwercman, A., & Giwercman, Y. L. (2011). Environmental factors and testicular function. Best Practice & Research.Clinical Endocrinology & Metabolism, 25(2), 391–402.
Guzick, D. S., Overstreet, J. W., Factor-Litvak, P., Brazil, C. K., Nakajima, S. T., Coutifaris, C., et al. (2001). Sperm morphology, motility, and concentration in fertile and infertile men. New England Journal of Medicine, 345(19), 1388–1393.
Han, J., Kamber, M., Pei, J. (2011). "Data Mining: Concepts and Techniques" The Morgan Kaufmann Series in Data Management Systems, 3rd edition, ISBN: 978-0123814791
Keel, B. A. (2006). Within-and between-subject variation in semen parameters in infertile men and normal semen donors. Fertility and Sterility, 85(1), 128–134.
Kolettis, P. N. (2003). Evaluation of the subfertile man. American Family Physician, 67(10), 2165–2172.
KNIME (2013). http://www.knime.org/ (May 2013)
Martini, A. C., Molina, R. I., Estofan, D., Senestrari, D., Fiol de Cuneo, M., & Ruiz, R. D. (2004). Effects of alcohol and cigarette consumption on human seminal quality. Fertility and Sterility, 82(2), 374–377.
R. Agarwal and R. Srikant (1994). Fast algorithms for mining association rules in large databases. Research Report RJ 9839, IBM Almaden Research Center, San Jose, California.
Rowe, P. J., & Comhaire, F. H. (2000). WHO manual for the standardized investigation, diagnosis and management of the infertile male. Cambridge University Press.
Slama, R., Eustache, F., Ducot, B., Jensen, T. K., Jorgensen, N., Horte, A., et al. (2002). Time to pregnancy and semen parameters: A cross-sectional study among fertile couples from four European cities. Human Reproduction, 17(2), 503–515.
Weka (2013). http://www.cs.waikato.ac.nz/ml/weka/ (May 2013)
WHO. (1999). WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction (4th ed.). Published on behalf of the World Health Organization by Cambridge University Press, Cambridge, UK.
Wong, W. Y., Zielhuis, G. A., Thomas, C. M., Merkus, H. M., & Steegers-Theunissen, R. P. (2003). New evidence of the influence of exogenous and endogenous factors on sperm count in man. European Journal of Obstetrics Gynecology and Reproductive Biology, 110(1), 49–54.
XLMiner (2013). (http://www.resample.com/xlminer/index.shtml (May 2013)

M. Abaidullah Anwar is working as Associate Professor and Deputy Dean of College of Engineering and Computing in Al Ghurair University, UAE. He received his Doctorate of Engineering with specialization in object-oriented databases from Kyushu Institute of Technology, JAPAN in 2001. Since 2001, he has been affiliated with renowned universities in GCC and Pakistan.

Naseer Ahmed is working as Director, Institutional Effectiveness and Planning at Al Ghurair University, UAE. He has received his PhD in 1991 from Heriot-Watt University, UK and has also a professional post-doctoral experience in the field of medical physics. His professional experience spans a number of academic assignments held at highly acclaimed institutions in South East Asia, South Asia, Canada, Saudi Arabia and UAE. Since last thirty years, he has been publishing his research work in a number of prestigious international journals. Currently his active areas of interests are quality enhancement, institutional effectiveness and planning, curriculum development, instructional methodology, and student assessment. In the past, Dr. Ahmed has been active member of various professional organizations and associations such as American Association of Physicists in Medicine, Institute of Physics, and American Physical Society. At present, Dr. Ahmed is an active member of Association of Institutional Research, USA.